Filtrage automatique de courriels : une approche adaptative et multi niveaux

Omar Nouali, Philippe Blache

Filtrage automatique de courriels une approche adaptative et multi niveaux

Omar NOUALI[*,**], Philippe BLACHE[**]

Résumé

Cet article propose un système de courriers électroniques paramétrable avec plusieurs niveaux de filtrage: un filtrage simple basé sur l'information contenue dans l'entête du courriel ; un filtrage booléen basé sur l'existence ou non de mots clés dans le corps du courriel ; un filtrage vectoriel basé sur le poids de contribution des mots clés du courriel ; un filtrage approfondi basé sur les propriétés linguistiques caractérisant la structure et le contenu du courriel.

Nous proposons une solution adaptative qui offre au système la possibilité d'apprendre à partir de données, de modifier ses connaissances et de s'adapter à l'évolution des intérêts de l'utilisateur et à la variation de la nature des courriels dans le temps. De plus, nous utilisons un réseau lexical permettant d'améliorer la représentation du courriel en prenant en considération l'aspect sémantique.

Mots clés : filtrage d'information, filtrage multi niveaux, apprentissage automatique, modèles linguistiques réduits.

COURRIEL AUTOMATIC FILTERING AN ADAPTIVE AND MULTI LEVELS APPROACH

Abstract

This article proposes an electronic mail system with several levels of filtering: a simple filtering based on contents analysis of the courriel header fields; a Boolean filtering based on the existence or not of key words in the body of the courriel; a Vectorial filtering based on weight of key words; a linguistic filtering based on the linguistic properties characterizing structure and contents of courriel.

We propose an adaptive solution by using an automatic learning method which allows the filtering system to learn from data, to modify its knowledge and to adapt to user's interests. Moreover, we use a lexical network to improve courriel representation and to take into account the semantic aspect.

Key words: information filtering, multi levels filtering, automatic learning, small scale linguistic models.

[*] Laboratoire de logiciels de base, CE.R.I.S.T., rue des 3 frères Aïssiou, Ben Aknoun, Alger, 16030, Algérie.
Email: onouali@mail.cerist.dz - Fax : 213 (21) 912126 - Tel. : 213 (21) 916211.
[**] LPL- Université de Provence, 29, Av. Robert Schuman, F-13621 Aix-en-Provence, France.
Email: pb@lpl.univ-aix.fr - Fax: +33 (0).42.59.50.96 - Tel: +33 (0) 42.95.36.23.

Sommaire

I. INTRODUCTION

Les moyens de communication électronique sont en rapide expansion, ils permettent facilement de créer et de diffuser de l'information auprès des utilisateurs. Aujourd'hui, le courrier électronique est le mode de communication le plus populaire. Il est devenu un moyen rapide et économique pour échanger des informations. Cependant, les utilisateurs d'Internet se retrouvent assez vite submergés de quantités astronomiques de courriels dont le traitement nécessite un temps considérable. Devant l'importance de ce phénomène, il est donc nécessaire aujourd'hui d'élaborer des outils efficaces capables de traiter et de filtrer le courriel.

La plupart des systèmes de filtrage de courriels existants enregistrent des lacunes ou faiblesses du point de vue efficacité de filtrage. Certains systèmes sont basés seulement sur le traitement de la partie structurée (par exemple, dans le cas de courriels non sollicités appelés *spam*, le filtrage opère généralement sur les adresses émettrices en se basant sur une liste noire des *spammeurs*), et d'autres sont basés sur un balayage superficiel de la partie texte du courriel en permettant aux utilisateurs d'écrire manuellement de règles logiques de filtrage à base de mots clés. La mesure de la pertinence repose uniquement sur la présence ou absence de mots clés dans le courriel traité.

Toute analyse effectuée sur ces bases ne peut que contenir une part d'imprécision. Par exemple, les courriels pertinents dont la représentation ne correspond qu'approximativement au profil ne seront pas sélectionnés par le système de filtrage. En effet, une représentation du contenu des courriels par des mots clés est particulièrement pauvre : les mots clés ne préservent qu'une faible fraction du sens.

Pour améliorer ces systèmes, notre motivation a été d'utiliser des ressources et des traitements linguistiques, d'une part, et d'explorer le potentiel des techniques d'apprentissage automatique, d'autre part.

Cet article propose un système paramétrable avec plusieurs niveaux de filtrage. Il offre à l'utilisateur la possibilité de choisir un niveau de filtrage et de basculer, à tout moment, d'un niveau à un autre selon son choix. Ces différents niveaux varient en terme de stratégie et de profondeur d'analyse. En effet, dans le cadre de ce travail, nous ne cherchons pas à faire une analyse complète et profonde du contenu des courriels, mais plutôt, une analyse partielle selon le niveau d'analyse choisi par l'utilisateur, qui permet de dégager des propriétés linguistiques qui devraient permettre de caractériser et de distinguer les différents types de courriels et de classer ensuite les nouveaux courriels. Nous proposons d'utiliser une méthode d'apprentissage automatique permettant au système d'apprendre à partir de données et de s'adapter à la nature des courriels dans le temps. Nous l'avons expérimenté sur un corpus de quelques types génériques de courriels bien particuliers : les courriels personnels, professionnels et les courriels indésirables (appelés *Spam*) qui continuent à polluer nos boîtes courriels de façon croissante. Nous présentons à la fin les résultats d'un ensemble d'expériences d'évaluation.

II. ETAT DE L'ART

Du fait que le domaine de filtrage de l'information électronique soit étroitement lié au domaine de la recherche d'information, les principales techniques actuelles employées dans le domaine du filtrage sont basées d'une façon directe ou indirecte sur les techniques des méthodes traditionnelles de recherche d'information [29]. Elles se basent sur l'occurrence d'un ensemble de mots clés pour identifier ou reconnaître les courriels pertinents (modèle booléen, modèle vectoriel…).

L'avantage de cette approche statistique repose principalement sur sa simplicité, mais elle est basée sur une hypothèse irréaliste qui est celle que tous les mots sont complètement indépendants. En effet, les systèmes les plus répandus, ayant participé à TREC (conférence de référence) ne prennent pas en compte l'ordre des mots, et les relations de dépendances existantes entre les éléments linguistiques (mots, syntagmes, chunks, phrases...). Par exemple, de tels systèmes (statistiques) ne font pas la différence entre «*le ministre de la culture*» et «*la culture du ministre*».

D'autres techniques tentent de capter le plus d'information sémantique (méthodes de traitement du langage naturel). Elles cherchent à améliorer les performances des systèmes de filtrage en unifiant la sémantique des textes et les profils des utilisateurs. Pour cela, nous avons besoin d'un modèle sémantique qui permet de représenter les intérêts de l'utilisateur, et de faire une compréhension du texte qui nécessite classiquement: une étape morphologique, lexicale, syntaxique, sémantique et même pragmatique [29].

Cette approche sémantique est intéressante et donne un filtrage efficace, mais difficile à appliquer à des textes tout venant couvrant des domaines variés. Elle est avantageuse dans un domaine spécifique et se complique rapidement quand il s'agit d'une généralisation. Elle nécessite de mobiliser d'importantes ressources linguistiques (dictionnaire) et des outils de traitement automatique du langage naturel (analyseur syntaxique, grammaire, représentation textuelle).

De ce fait et vu la diversification des tendances classiques et actuelles, il n'y a pas encore de conclusion concrète. Indexer des phrases ou expressions au lieu de mots clés apporte des améliorations certaines à l'efficacité du filtrage au frais d'un prétraitement élaboré (analyse partielle ou totale et analyse syntaxique de la phrase). L'état actuel des connaissances ne permettent pas de concevoir un système capable de comprendre n'importe quelle phrase écrite en langage naturel et de fonctionner dans n'importe quel contexte d'utilisation, sans aucune adaptation ou modification éventuelle. La quête pour un analyseur robuste capable d'analyser des textes libres en profondeur, a mené au développement de toute une gamme d'analyseurs ces dernières années. Ces analyseurs varient en terme de stratégie : le déterminisme contre le non déterminisme, l'analyse partielle contre l'analyse complète. Cependant, aucun d'entre eux n'est capable de faire une analyse complète et profonde du contenu des textes libres.

Dans ce contexte, nous proposons une analyse partielle du contenu des courriels utilisant plusieurs niveaux d'analyse qui permettent de dégager des propriétés linguistiques. Ces propriétés portant sur la structure et le contenu des courriels devraient permettre de distinguer les différents types de courriels.

Notre motivation a été d'utiliser les techniques d'apprentissage automatique. En effet, la croissance accélérée d'Internet et la grande quantité d'information devenue disponible sur ce réseau, ces dernières années, ont motivé les recherches dans ce domaine. Une grande majorité des travaux utilise la cooccurrence lexicale comme base de leur analyse [3, 10, 30, 32, 33]. Nous citons les méthodes basées sur une approche probabiliste [33], les méthodes à base de règles de décision [4, 21], à base de régression (méthode Rocchio, TFIDF...) [19], les arbres de décisions [33], le k-proche voisin (kNN) [34], Naive Bayes [21, 23]. Enfin, les réseaux de neurones [2] et les machines à vecteur support (SVM) [20]. En plus des approches de sémantique lexicale qui font recours à des ressources lexicales (thésaurus, dictionnaires...) tel que Wordnet [21, 27], d'autres études ont également été menées en vue d'exploiter des informations de plus haut niveau que le mot. Nous citons les travaux qui intègrent des séquences de mots en accord avec une grammaire [22] ou purement statistiques [2, 9].

III. ARCHITECTURE DU SYSTEME

L'architecture globale du système est constituée principalement des modules suivants (figure 1) :

– un module de prétraitement qui détermine la langue de chaque courriel et le prépare aux différentes étapes ultérieures de l'analyse en sélectionnant les connaissances nécessaires selon le niveau de filtrage choisi par l'utilisateur ;

– un analyseur qui a pour but d'identifier les informations pertinentes permettant de caractériser et de représenter le contenu des courriels. Il utilise un ensemble de connaissances de base et délivre en sortie une représentation vectorielle associée, selon le niveau de filtrage choisi ;

– un module d'expansion permettant d'améliorer la représentation du courriel ;

– un réseau de neurones qui modélise les différents profils de l'utilisateur. Il constitue la connaissance du système de filtrage. Il permet de comparer un nouveau courriel avec les différents profils de l'utilisateur;

– un système expert qui modélise les différentes actions de filtrage telles que supprimer, sauvegarder, signaler...

– un module d'apprentissage qui permet d'améliorer l'efficacité et les performances du système.

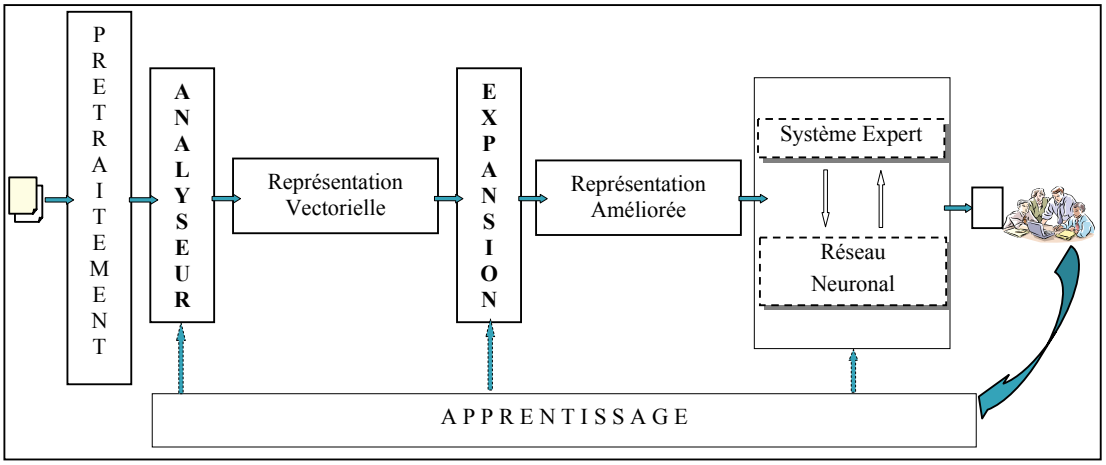


FIG. 1- Architecture du système.

Architecture of the system.

IV. MODELE UTILISATEUR

Le système de filtrage présente à l'utilisateur des types ou classes de courriels clairement identifiables pour lui sous forme d'une typologie générale. En effet, l'existence d'une typologie des courriels faisant défaut, nous nous somme limités à quelques types génériques de courriels bien particuliers : les courriels personnels, professionnels et les courriels indésirables (appelés *Spam*) qui continuent à polluer nos boîtes courriels de façon croissante.

Les courriels personnels regroupent tous les courriels familiaux, ceux provenant d'amis, ainsi que les courriels personnels-professionnels (collègue -collègue, étudiant-professeur...).

Les courriels professionnels regroupent les appels à communication, les annonces de livres, les articles, les courriels de direction...

Enfin, les courriels non sollicités appelés Spam. Il s'agit de courriels publicitaires proposant des services, des produits miraculeux (maigrir en un temps record...), offres de voyages à prix attractif, opportunités d'investissement pour devenir riche en peu de temps, propositions de cartes de crédit à taux d'intérêt réduit, courriels pornographiques...

Cette typologie constitue donc une sorte de profil de base, qui permettra d'aider l'utilisateur à décrire et élargir ses propres profils (classes):

- soit utiliser ou combiner des types prédéfinis ;
- soit proposer de nouveaux types, introduits dans le système sous formes différentes :

*** sous forme de mots clés**: l'utilisateur introduit une liste de mots clés et pour chaque mot clé, il associe un poids qui représente son degré d'importance. Ce type de profils est amélioré et augmenté de propriétés linguistiques par un apprentissage au fur et à mesure des nouveaux courriels entrants ;

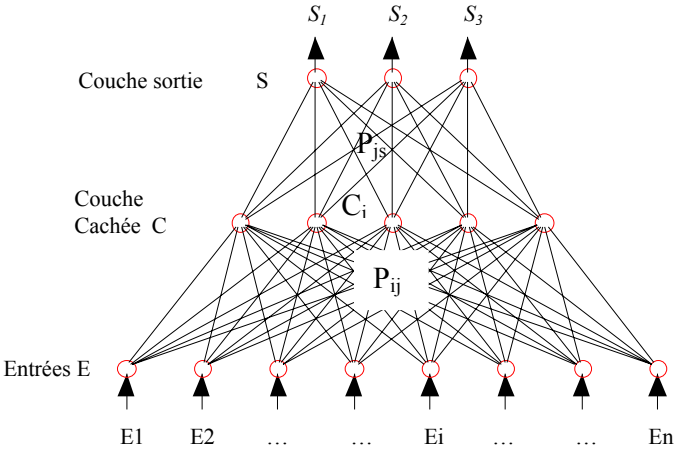
*** sous forme de texte:** ici, l'utilisateur introduit des textes et c'est au système de se charger de l'extraction des mots clés et des propriétés linguistiques et de leur attribuer un poids ;

***sous forme d'url** (ex : adresse d'un serveur).

Chaque forme subira un traitement spécifique pour être représentée dans le système.

IV.1. Réseau de neurones

Le modèle adopté pour notre système est un réseau de neurones non récurrents (absence de boucles) à trois couches (Figure 2). Une couche en entrée qui reçoit les entrées du réseau *(E)*. Une couche cachée représentant l'ensemble des connaissances *(C)*. Une couche de sortie qui représente les types de courriels *(S)*.



S_j : Neurone de sortie d'indice *j*.
C_j : Neurone caché d'indice *j*.
E_i : Neurone en entrée d'indice *i*.
P_{ij} : Valeur du poids de la connexion du neurone d'indice *i* de la couche d'entrée vers le neurone d'indice *j* de la couche cachée.
P_{js} : Valeur du poids de la connexion du neurone d'indice *j* de la couche cachée vers le neurone d'indice *s* de la couche de sortie.

FIG. 2- Architecture d'un réseau à trois couches.

Neural Network Architecture with three layers.

IV.2. Apprentissage

Dans un réseau de neurones, la connaissance est codée par la valeur des poids des différentes connexions. Ce codage est estimé par apprentissage. Nous appelons apprentissage, la procédure qui consiste à estimer les paramètres d'un système, afin que celui-ci remplisse au mieux la tâche qui lui est affectée. Il s'agit, à partir d'un ensemble d'exemples observés, d'induire une procédure ou une règle (classification dans notre cas). La procédure générée devra fournir correctement les réponses appropriées aux exemples de l'échantillon mais surtout avoir un bon pouvoir prédicatif pour répondre correctement aux nouvelles descriptions (généralisation).

Entrée : un échantillon de courriels;
un réseau avec une couche d'entrée *E,* une couche cachée *C*,
une couche de sortie *S*, 3 cellules, une fonction d'activation : Sigmoïde [14]
Initialisation à 0.5 des poids p_i des connexions entre neurones et le *pas* à *0.1* [16].
Répéter
Prendre un exemple *x* de *E* et calculer *S*
- - calcul des erreurs *DS* et *DC* par rétropropagation
Pour toute cellule de sortie *i* $DS_i = S_i*(1-S_i)*(1-S_i)$ **finPour**
Pour chaque cellule j de la couche cachée
$DC_j = C_j * (1 - C_j) \sum_{k \in Succ(j)} DS_k * P_{ks}$
finPour
- - mise à jour des poids
Pour tout poids $P_{ij}(t+1) = P_{ij}(t) + r*DC_j*S(x_{ij})$ **finPour**
Pour tout poids $P_{js}(t+1) = P_{js}(t) + r*DS_i*S(C_j)$ **finPour**
r : le taux d'apprentissage ; *t* : le numéro du cycle.
finRépéter

FIG. 3- Algorithme d'apprentissage.

Learning algorithm.

Nous avons travaillé avec un corpus de 1200 courriels dit « base d'apprentissage » composé de courriels *spam*, *personnel* et *professionnel*.

Le réseau est entraîné sur cette base d'apprentissage, dans le but de correctement catégoriser un nouveau courriel, par l'algorithme de propagation arrière ou rétro- propagation (Figure 3) qui consiste à corriger les poids des connexions en fonction des erreurs commises. La correction se fait de la couche de sortie à la couche d'entrée.

V. NIVEAUX DE FILTRAGE

Notre système de filtrage est destiné à traiter des informations textuelles semi structurées, il s'agit de courriels:

- une entête bien définie (Header) : *from, to, subject...*
- un corps libre non structuré.

Il est paramétrable avec plusieurs niveaux de filtrage :

- filtrage basé sur l'information structurée (entête du courriel) ;
- filtrage basé sur le contenu (texte) : superficiel (ou booléen), intermédiaire (ou vectoriel) et approfondi (ou linguistique).

V.1. Filtrage basé sur l'information structurée

C'est un filtrage simple, qui offre à l'utilisateur la possibilité de définir un ensemble de règles de filtrage sur l'information contenu dans l'ensemble de l'entête du courriel. Il utilise un système expert (Figure 4).

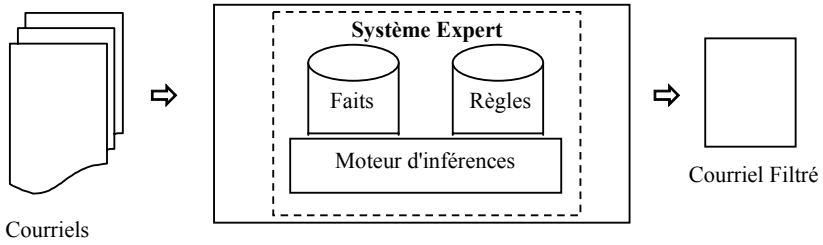


FIG. 4- Filtrage par Système Expert.

Filtering by expert system.

L'avantage d'utiliser un système expert est qu'il peut justifier les décisions concernant l'opération de filtrage auprès de l'utilisateur en lui présentant l'ensemble des règles ayant contribué à la prise de chaque décision. Ce qui permet à l'utilisateur, à tout moment, de mettre à jour la base de règles.

Les connaissances du système expert sont organisées sous forme de règles de production de la forme:

***IF** (suite de conditions) **THEN** (suite d'actions)*

Les conditions portent sur les différents champs du courrier (*From, Subject, to, date*,…) et les actions à entreprendre sont: sauvegarder ou classer, déliter, générer une réponse automatique...

Exemples:

- règle 1: *IF From= 'omar@tassili.cerist.dz' THEN delete (courriel);*

si le courriel provient de la personne *Omar*, alors effacer le courriel.

- règle 2: *IF Subject Contains 'sport' THEN save (courriel) in sport folder;*
si le mot *'sport'* apparaît dans le champ *subject* du courriel alors sauvegarder le courriel dans le répertoire *'sport'*.

V.2. Filtrage basé sur le contenu (texte)

Ce type de filtrage nécessite une étape de prétraitement qui consiste à identifier la langue de chaque courriel. L'identification de la langue est une étape nécessaire : elle permet de caractériser la langue du courriel (Français, Anglais). Elle utilise des anti-dictionnaires propres à chaque langue (mots outils : articles, prépositions...). La prise en compte de nouvelles langues est simple (ajouter un anti-dictionnaire propre à chaque nouvelle langue). Puis, une normalisation des mots est effectuée en réduisant les variantes morphologiques à une forme commune (souvent appelé terme). Pour cela, nous utilisons un analyseur morphologique flexionnel, analyseur FLEMM [28], fourni par l'ATILF (Analyse et Traitement Informatique de la Langue Française).

La figure 5 présente un exemple de résultat de lemmatisation.

Message : …*C'est le moment aussi pour prendre des contacts…*

Message lemmatisé : C'/PRV:3p:_:s:n/ce est/ECJ:3p:s:pst:ind/être:3g le/DTN:m:s/le moment/SBC:_:s/moment aussi/ADV/aussi pour/PREP/pour prendre/VNCFF/prendre des/DTC:_:p/du contacts/SBC:_:p/contact ./.

Jeu d'étiquettes : ***DTN :*** *Déterminant.* ***PRV :*** *Pronom « supporté » par le verbe (conjoint, clitique).* ***ECJ :*** *Verbe « être », conjugué.* ***ADV :*** *Adverbe.* ***SBC :*** *Substantif, nom commun.* ***VNCFF :*** *autre Verbe, non conjugué, infinitif.* ***PREP :*** *Préposition.* ***DTC :*** *Déterminant de groupe nominal, contracté.* ***Sg:*** *Singulier...*

Traits : *personne (1p, 2p, 3p), genre (m : masculin, f : féminin), nombre (s : singulier, p : pluriel), temps (pst : présent, impft : imparfait, fut : futur, ps : passé simple), mode (ind : indicatif, subj : subjonctif, cond : conditionnel, imper : imperatif), groupe (1gr, 2gr, 3gr), cas (n : nominatif, a : accusatif, d : datif et o : oblique).*

FIG. 5- Résultat du lemmatiseur.

Result of Streaming operation.

Nous distinguons 3 types de filtrage basé sur le contenu : superficiel, intermédiaire et approfondi.

V.2.1. Filtrage superficiel

C'est un filtrage booléen, qui traite le contenu du courriel, mais d'une façon très superficielle. Il est basé sur la comparaison exacte entre le profil et les courriels. Il est basé sur l'existence ou non de mots clés dans le corps du courriel. L'utilisateur exprime ses profils par des mots qui doivent exister ou ne doivent pas exister dans le courriel à recevoir.

Le système sélectionne les courriels qui satisfont une expression logique sur les termes du profil. Les opérations de base pour ce modèle sont les connecteurs logiques : *ET (AND), OU (OR)* et *SAUF (NOT)*. Par exemple, le profil exprimé par l'expression logique *P= (« intelligence » OU « raisonnement ») ET « artificiel »*, permet de sélectionner tous les courriels contenant le terme *« artificiel »* et un des termes *« intelligence* » ou *« raisonnement »*.

V.2.2. Filtrage intermédiaire

C'est un filtrage vectoriel, basé sur le poids de contribution des différents termes du courriel. Il utilise comme support de base un modèle utilisateur représentant les différents profils. Chaque profil est représenté par un ensemble d'entités lexicales les plus pertinentes. Le profil est soit généré automatiquement à partir d'un ensemble de courriels, soit introduit par l'utilisateur sous forme de mots clés.

V.2.3. Filtrage approfondi

Le filtrage est basé sur des indices qui portent sur la structure et le contenu des courriels. Le courriel passe par plusieurs niveaux d'analyse (Figure 6) : l'objectif de chaque niveau est d'analyser le courriel et d'en extraire un ensemble de propriétés sous forme d'indices. L'ensemble des propriétés, extraites au fur et à mesure de l'analyse, constitue la représentation interne du courriel. Elle est donc crée dynamiquement à chaque récupération d'un nouveau courriel.

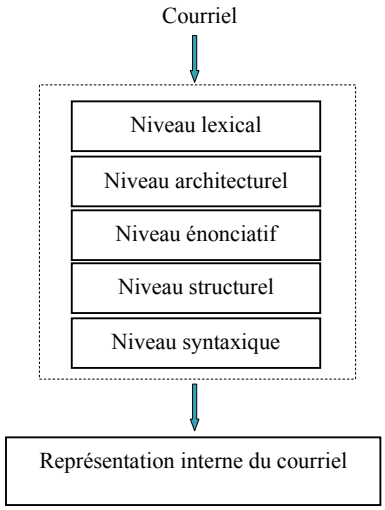


FIG. 6- Niveaux d'analyse.

Analysis levels.

Le filtrage approfondi nécessite un ensemble de connaissances linguistiques de base pour analyser, extraire les différentes propriétés et construire la représentation interne de chaque courriel. Pour cela, nous avons défini et identifié d'une façon automatique un ensemble de connaissances linguistiques que nous avons classé en plusieurs niveaux (Tableau I) : lexical, architectural, énonciatif, structurel et syntaxique.

Ces connaissances sont indépendantes du domaine d'application. Il s'agit d'un ensemble d'indicateurs sur le courriel, automatisables, qui permettent de caractériser un courriel et de le situer par rapport aux autres. Ils permettent de rapprocher les courriels qui appartiennent à la même classe ou éloigner ceux qui appartiennent à des classes différentes.

Table I.- Connaissances linguistiques.

Linguistic Knowledge.

Connaissances lexicales	***Mots simples***	*money, price, product, credit, order, opportunity, marketing, investment, advertisement, sex, travel... bisous, bonjour, famille, galère, maman, médecin, ok, papa, plaisir, samedi, vacances, visa, voiture... cher, collègue, communication, conférence, cordial, critères, date, madame, monsieur, salutation...*
	Mots composés	*credit card, free investment, half price, home business, live sex, low price, phone number, special bonus...à bientôt, après-midi, parce que, rendez-vous, week-end... appel à communication, comité de lecture, critères de sélection, date de soumission...*
Connaissances architecturales		*Titre, section, introduction, conclusion, Image, dessin, Ponctuation, type du courriel, fichier attaché, nombre de destinataires, domaine des adresses, longueur de l'entête, longueur du courriel, abréviation, horaire d'envoi...*
Connaissances énonciatives		*1ere pers (singulier, pluriel), 2eme pers (singulier, pluriel),3 pers (singulier, pluriel),déterminants (mon, ton, son, ce...), retours de courriers (réponses), énoncer, admettre, dire, déclarer, remarquer, protester, penser, croire, révéler, supposer, estimer...*
Connaissances structurelles		*Addition (à cela s'ajoute qu, ainsi qu, aussi, d'autre part, de plus...), Analogie (c'est-à-dire, comme, de la même façon, de même...), But (pour qu, de sorte qu...), Cause (afin qu, c'est pourquoi...), Exemple (à savoir, par exemple...)...*
Connaissances syntaxiques		*Taux de pronoms, taux de déterminants, nominalisation, infinitif, participe passé, coordination, négation, subordination, interrogation, interjection, forme active, forme passive...*

Les connaissances lexicales sont divisées en deux types : un vocabulaire de mots simples *(MS)* et un vocabulaire de mots composés ou phrases très courtes *(MC)*.

La construction du vocabulaire *MS* suit les étapes suivantes :
(1) : élimination des mots outils ;
(2) : normalisation des termes ;
(3) : réduction du vocabulaire.

Le vocabulaire MC est généré à partir des listes *bigrammes* et *trigrammes* apprises par le système.

Le critère utilisé pour réduire le vocabulaire est la mesure de l'information mutuelle [34]. Cette mesure numérique permet de déceler les mots qui « s'attirent », c'est-à-dire qui tendent à apparaître en même temps. En effet, l'information mutuelle MI (*t, C*) mesure la dépendance d'un terme *t* et d'une classe *C*. Elle est définie par :

$$MI(t,C) = \log_2 \frac{p(t,C)}{p(t)p(C)} \approx \frac{N * p}{(p+p')(p+q)}$$

Avec :

p(t, C) : probabilité conjointe d'apparition de *t* et *C ;*
p(t) : probabilité à priori du terme t ;
p(C) : probabilité à priori de la classe C ;
p *:* nombre de courriels de classe *C* qui contiennent le terme *t ;*
q *:* nombre de courriels qui ne sont pas de classe *C* mais qui contiennent le terme *t ;*
p' *:* nombre de courriels de classe *C* qui ne contiennent pas le terme *t* ;
q' : nombre de courriels qui ne sont pas de classe *C* et qui ne contiennent pas le terme *t ;*
N : nombre total de courriels du corpus.

Une information mutuelle élevée entre un terme et une classe est le signe d'un lien fort entre ces deux éléments.

L'architecture d'un texte donné est définie par un ensemble d'objets textuels représentant les différentes zones textuelles (entêtes, titres, corps, paragraphes, sections, listes, tableaux, chapitres...) et des relations entre ces objets (inclusion, liens sémantiques...). L'identification de ces objets est basée sur la ponctuation des textes (la ponctuation de la phrase qui inclut le format de titres, la forme des paragraphes, les notes de bas de page...).

La connaissance énonciative concerne le locuteur ou l'énonciateur. Elle est représentée dans l'énoncé par l'intermédiaire d'indicateurs linguistiques : pronoms personnels, formes verbales, formes temporelles...

La connaissance structurelle est représentée par un certain nombre de marqueurs linguistiques (un ensemble de mots clés) précisent la relation explicite ou un ensemble de relations potentielles (causalité, but...) entre deux segments de textes reliés par ce marqueur.

Pour l'aspect syntaxique, nous ne cherchons pas à retrouver précisément la structure syntaxique de chaque énoncé, cependant nous cherchons à identifier un ensemble d'indices syntaxiques (taux de pronoms, forme active…) pour alimenter notre système de filtrage de courriels.

L'extraction de certains indices (d'ordre syntaxique) nécessite une phase d'étiquetage préalable. En effet, nous avons utilisé un étiqueteur morphosyntaxique, analyseur de Brill [8]. La figure 7 présente un exemple de résultat de l'étiqueteur.

Message : … *C'est le moment aussi pour prendre des contacts…*

Message étiqueté : *C'/PRV:sg est/ECJ:sg le/DTN:sg moment/SBC:sg aussi/ADV pour/PREP prendre/VNCFF des/DTC:pl contacts/SBC:pl ./.*

Jeu d'étiquettes : ***DTN :*** *Déterminant.* ***PRV :*** *Pronom « supporté » par le verbe (conjoint, clitique).* ***ECJ :*** *Verbe « être », conjugué.* ***ADV :*** *Adverbe.* ***SBC :*** *Substantif, nom commun.* ***VNCFF :*** *autre Verbe, non conjugué, infinitif.* ***PREP :*** *Préposition.* ***DTC :*** *Déterminant de groupe nominal, contracté.* ***Sg:*** *Singulier...*

FIG. 7- Résultat de l'étiqueteur Brill.

Result of Brill tagger.

La sortie de l'analyse constitue l'entrée du processus sémantique qui consiste à compléter et à améliorer la représentation interne de chaque courriel. Ce traitement sémantique utilise un réseau lexical de base construit par apprentissage.

La figure 8 présente un exemple simple de résultat d'analyse.

Message source :

. . .
> Bonjour Alain,
> J'espère que ton déplacement en suisse a été bénéfique.
> Raconte moi un peu ton aventure.
> Omar
Bonjour Omar.
Tout s'est très bien passé. C'est dommage que tu n'aies pas pu venir,
car c'est le moment d'apprendre beaucoup de choses.
C'est le moment aussi pour prendre des contacts.
Bon courage.
Alain.
p.s.: envoie moi l'article.

Message après analyse :

…
> Bonjour Alain,
> J'espère que ton déplacement en suisse a été bénéfique.
> Raconte moi un peu ton aventure.
> Omar
Bonjour Omar.
Tout s'est très bien passé. C'est dommage que tu n'aies pas pu venir,
car c'est le moment d'apprendre beaucoup de choses.
C'est le moment aussi pour prendre des contacts.
Bon courage.
Alain.
p.s.: envoie moi l'article.

Pour ne pas surcharger le texte, certaines propriétés ne sont pas représentées.

Jeu de propriétés :
- Propriétés lexicales
Bonjour, C'est, dommage, moment, beaucoup, Bon courage
- Propriétés Matérielles:
Type du message : txt, p.s., Langue: Français, Destinataires : 1, Auteur : regnier@lpl.univ-aix.fr, Taille du message (mots):38, Taille du message (phrases):7, Horaire : jour.
- Propriétés énonciatives:
Prem. Pers. Sing., moi, Deux. Pers. Sing., tu, Réponse, >.
- Propriétés Syntaxiques:
Pronoms, Substantifs, Interjections, Déterminants, Adverbes, Adjectifs, Infinitifs, Participes Passés, Coordinations, Subordinations, Abréviations.
- Propriétés structurelles:
Addition.

FIG. 8- Résultat de l'analyse linguistique.

Result of linguistic analysis.

VI. EXPANSION SEMANTIQUE

L'idée de base est qu'en général, les concepts définis pour représenter les intérêts des utilisateurs ne sont pas forcément les mêmes que ceux extraits à partir des courriels (le langage naturel est très riche). Pour cela, nous proposons d'utiliser un réseau lexical permettant d'améliorer la représentation du courriel en prenant en considération les termes qui existent dans le courriel et qui n'existent pas dans le profil. Il s'agit de les remplacer par des termes du profil sémantiquement proches. C'est-à-dire, impliquer d'autres mots (proches) dans la décision du filtrage, même s'ils n'apparaissent pas explicitement dans le courriel. Le réseau permet de relier des termes même s'il n'existe pas de lien visible entre ces termes. En effet, deux termes sont utilisés dans des contextes similaires, ils vont avoir des représentations similaires ou proches.

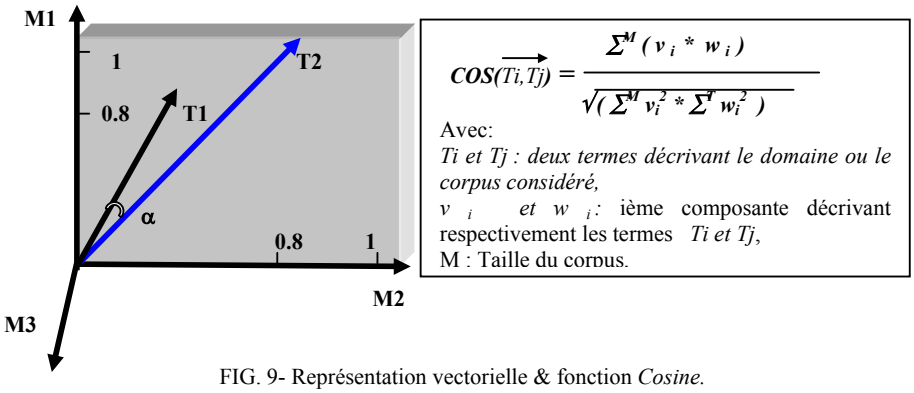
VI.1. Construction du réseau lexical

La construction du réseau lexical est basée sur une idée simple qui consiste à :

1. constituer un corpus de courriels pour un profil donné ;
2. construire les vecteurs courriels dans l'espace des termes ;
3. inverser ces vecteurs courriels pour construire d'autres vecteurs qui représentent l'ensemble de termes dans un espace dont les axes sont les courriels (vecteurs termes);
4. calculer la similarités des termes deux à deux ;
5. regrouper ensemble les termes les plus proches.

Cette connaissance est construite initialement sur la base d'un corpus de courriels qui sera enrichie et adaptée aux différents utilisateurs progressivement au fur et à mesure de l'utilisation du système par ces derniers.

Le calcul de la similarité entre termes se mesure à l'aide de la formule *Cosine* qui calcule le cosinus de l'angle entre leurs vecteurs respectifs (Figure 9).



$$\overrightarrow{COS(Ti,Tj)} = \frac{\Sigma^{M}(v_i * w_i)}{\sqrt{(\Sigma^{M} v_i^2 * \Sigma^{T} w_i^2)}}$$

Avec:
Ti et Tj : deux termes décrivant le domaine ou le corpus considéré,
v_i et w_i : ième composante décrivant respectivement les termes *Ti et Tj*,
M : Taille du corpus.

FIG. 9- Représentation vectorielle & fonction *Cosine.*

Vectorial representation & Cosine function.

Plus le cosinus de l'angle entre les deux vecteurs est proche de 1, plus les vecteurs sont proches ce qui implique une plus grande ressemblance entre les deux termes. La figure 10 donne un aperçu du réseau généré automatiquement par le système (cas du domaine *SPAM*):

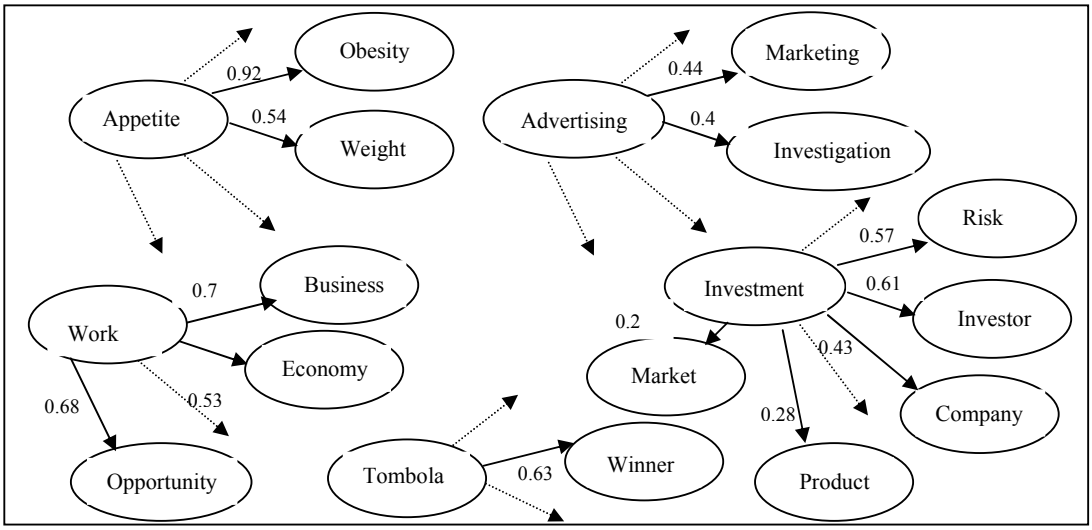


FIG. 10- Réseau lexical.

Lexical network.

L'exploration du réseau lexical permet au système d'améliorer la représentation d'un courriel contenant, par exemple, le terme *« obesity »* (terme non existant dans le vocabulaire lexical du système) en le remplaçant par sa classe, c'est-à-dire le terme « *appetite* ».

VI.2. Apprentissage assisté

L'apprentissage est un aspect intrinsèque de l'intelligence et une nécessité pour s'adapter à un environnement évolutif. Il permet d'améliorer l'efficacité et les performances d'un système.

Le système dispose d'un **apprentissage assisté** appelé ***feed-back*** où l'utilisateur est invité à donner son avis sur le comportement du système, ce qui lui permet d'approcher la pertinence de l'utilisateur et de s'adapter ainsi à ses besoins. L'apprentissage agit sur le réseau lexical, qui consiste à modifier les liens du réseau (terme-terme). L'utilisateur peut aussi ajouter et supprimer des mots à sa demande. Un agent est chargé de surveiller le comportement de toutes les structures sémantiques qui contribuent à la conception du nouveau vecteur courriel dans l'espace du profil lors de la session de filtrage. Il sauvegarde une trace de chaque substitution réalisée. Ensuite, après validation des résultats de filtrage par l'utilisateur, toutes les structures sémantiques ayant contribué à la bonne décision seront renforcées (augmenter le degré de ressemblance) et celles ayant contribué à l'échec du système seront pénalisées (diminuer le degré de ressemblance). Un tel traitement ne donne de bons résultats que s'il se fait régulièrement. En effet, après une certaine période d'utilisation ou d'apprentissage, le système de filtrage sera doté d'un réseau lexical efficace et plus spécifique à l'utilisateur.
De même, nous pouvons rendre le réseau intelligent capable d'évoluer et de supprimer les structures fausses en lui appliquant les algorithmes génétiques [29].

VII. EVALUATION

Nous avons mené des tests pour mesurer les performances des différents nivaux de filtrage du système et montrer comment le réseau lexical ainsi que l'opération d'apprentissage agissent sur l'efficacité du filtrage.

Pour mesurer les performances nous utilisons les mesures suivantes :

$$rappel = \frac{\alpha}{\alpha+\gamma} \; ; \quad précision = \frac{\alpha}{\alpha+\beta} \; ; \quad p\_globale = \frac{\alpha+\delta}{\alpha+\beta+\gamma+\delta}$$

Où :

α : courriels de classe ou type *C* (s*pam, personnel ou professionnel)* correctement filtrés (classés) par le système ;

β : courriels n'appartenant pas à la classe *C*, incorrectement filtrés par le système ;

γ : courriels de classe *C*, incorrectement non filtrés (rejetés) par le système ;

δ : courriels n'appartenant pas à la classe *C*, correctement non filtrés par le système ;

p_globale (précision globale) : rapport entre le nombre total de courriels correctement filtrés et correctement non filtrés et le nombre total de courriels.

VII.1. Filtrage simple

Nous mesurons les performances du système en considérant un modèle de base constitué uniquement d'un ensemble limité de règles de filtrage simples (Figure 11). Les règles portent sur l'information contenue dans les différents champs de l'entête.

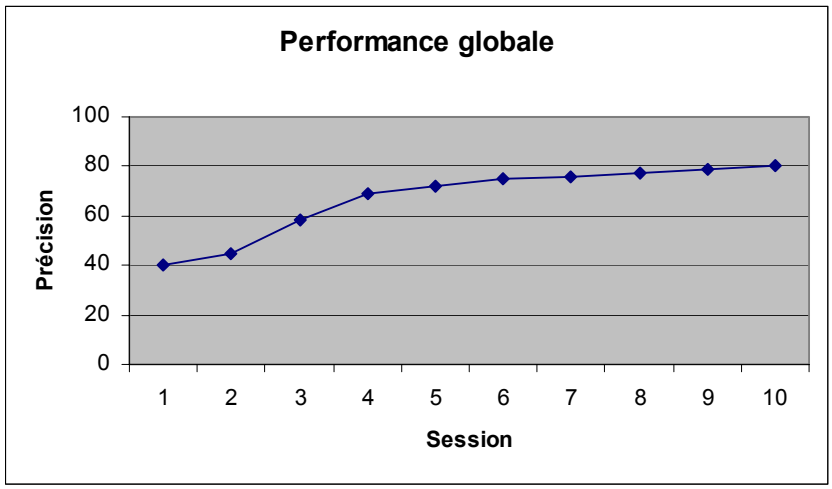


FIG. 11- Performances du filtrage simple.

Simple Filtering Performance.

Au début de l'expérience, nous remarquons un taux d'erreur élevé, et un écart significatif entre le taux de *précision* et le taux de *rappel* (système considère certains *spam* comme des courriels légitimes...). En effet, les règles définies initialement par l'utilisateur ne couvrent pas la totalité des courriels considérés. Ce qui explique la difficulté rencontrée par l'utilisateur pour décrire ses règles. Après plusieurs sessions, nous constatons que le système s'améliore progressivement au fur et à mesure de son utilisation. En effet, l'utilisation d'un système expert a permis à l'utilisateur d'afficher et de mettre à jour les règles ayant contribué à l'opération de filtrage. A la fin de l'expérience, nous constations que ce type de filtrage est efficace lorsqu'il y a presque autant de règles que de courriels (chaque courriel lui correspond une règle spécifique)

VII.2. Filtrage intermédiaire

Initialement, nous considérons un modèle de base (MB) constitué uniquement de mots simples. Ensuite, nous lui ajoutons un ensemble de mots composés (MC).

Table II.- Performances en fonction des mots composés.

Performance using phrases.

Caractéristiques	**Performance globale**		
	Personnel	*Professionnel*	*Spam*
MB	83%	88%	87,7%
MB + MC	83%	89%	87%
MB + MC +Pondération	83%	91%	91,4%

Nous ne constatons pas une amélioration des performances (Tableau II). En effet, les mots composés corrèlent avec les types de courriels considérés, mais statistiquement sont insignifiants (valeur faible). Ensuite, nous avons donc modifié l'importance de ces différents mots composés, en leur attribuant une forte valeur du poids. Les résultats des tests étaient nettement meilleurs (ex : 91% pour le profil *spam*).

VII.3. Filtrage approfondi

Dans un premier temps, nous considérons un modèle lexical constitué de mots simples et de mots composés (ML). Ensuite, nous lui ajoutons d'autres propriétés linguistiques (PL) de type architectural, énonciatif, structurel et syntaxique. Nous considérons que les propriétés dont le nombre d'occurrences dépasse un certain seuil fixé.

Les résultats sont donnés dans le Tableau III.

Table III.- Performances avec modèle linguistique.

Performance using linguistic model.

Catégories	ML	ML + PL
Personnel	85%	92%
Professionnel	91%	93%
Spam	90%	95%

Nous constatons que les performances globales du système sont améliorées. Ceci s'explique par le fait que les courriels rejetés par le système par absence de mots clés ou valeur très faible, sont acceptés cette fois ci, et ceci à cause de la présence de certaines propriétés linguistiques (PL). Par exemple, les courriels personnels sont caractérisés par l'utilisation de pronoms personnels (1ère et 2ème personne du singulier).

VII.4. Apprentissage assisté ou feedback

L'expérience consiste à présenter au système un ensemble de courriers à filtrer en plusieurs sessions. Puis mesurer à chaque fois la précision et le rappel et effectuer un apprentissage assisté pour mesurer son efficacité et son influence sur les deux facteurs.

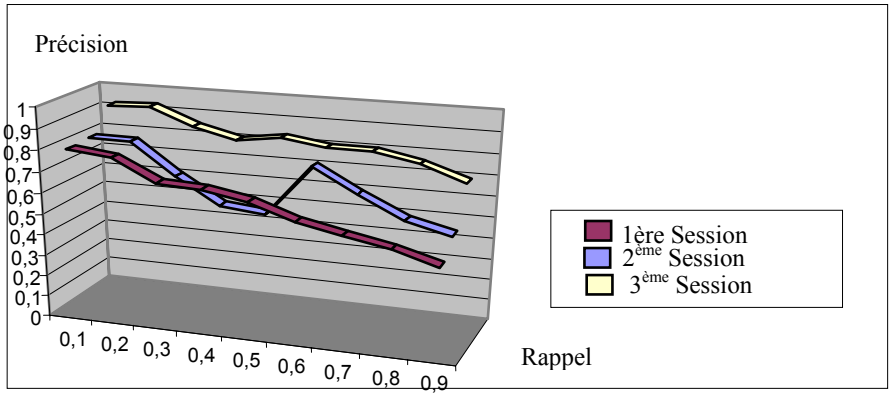


FIG. 12- Apprentissage assisté.

Feedback.

Après plusieurs sessions d'apprentissage assisté, nous constatons que les performances du système sont améliorées. En effet, dans le filtrage avec réseau lexical, la cooccurrence des termes est prise en considération, ce qui permet d'augmenter le taux de *rappel* tout en gardant une bonne précision. C'est-à-dire, la nouvelle représentation obtenue de certains courriels permet au système de les bien filtrer même s'ils ne partagent pas de termes communs avec les profils. Par exemple, en ce qui concerne les

courriels de type *personnel*, on obtient un taux de filtrage avec 92% de précision pour un rappel de 70%.

Le réseau lexical nécessite donc plusieurs sessions d'apprentissage assisté pour améliorer la qualité de ses résultats. Il est donc nécessaire de lancer l'apprentissage *feedback* régulièrement, par exemple après chaque session de filtrage.

VIII. CONCLUSION

Cet article propose un système de courriers électroniques paramétrable avec plusieurs niveaux de filtrage:

(1) un filtrage simple basé sur l'information structurée, qui offre à l'utilisateur la possibilité de définir un ensemble de règles de filtrage sur l'information contenu dans l'ensemble de l'entête du courriel ;

(2) un filtrage booléen appelé superficiel basé sur l'existence ou non de mots clés dans le corps du courriel ;

(3) un filtrage vectoriel appelé intermédiaire, basé sur le poids de contribution des mots clés du courriel ;

(4) un filtrage approfondi basé sur les propriétés linguistiques caractérisant le contenu ainsi que l'utilisation d'un réseau lexical permettant d'améliorer la représentation du courriel en prenant en considération l'aspect sémantique. Ces propriétés constituent un ensemble d'indices qui portent généralement sur la structure et le contenu des courriels.

Pour la mise en pratique du modèle de connaissances, une typologie des courriels est présentée à l'utilisateur lui permettant de l'aider dans la tâche de création de ses propres profils.

A travers les différentes expériences réalisées, nous avons montré l'applicabilité et l'adaptabilité d'une approche adaptative et multi niveaux au processus de filtrage. Elle permet au système d'apprendre et d'améliorer ses connaissances, et ceci selon le niveau de filtrage choisi par l'utilisateur.

Contrairement aux systèmes existants, notre système fait appel à des propriétés linguistiques (matérielles, énonciatives, structurelles et syntaxiques) qui permettent d'aider à améliorer les résultats de filtrage (filtrage approfondi). En effet, ces propriétés permettent de relier un message à un profil même s'ils n'ont pas de mots clés en commun (propriétés lexicales).

Pour l'aspect sémantique, notre approche fait recours à un réseau lexical : il regroupe les mots sémantiquement proches et permet d'améliorer la représentation de courriels à filtrer et d'augmenter donc les chances d'apparier un courriel et un profil. En effet, les termes du courriel sont automatiquement propagés en suivant les liens exprimés dans le réseau, de manière à disposer d'une description plus étendue de ce courriel. Ce traitement d'équivalence sémantique permet d'augmenter les performances du système.

Notre système est complètement indépendant du domaine de connaissances. Il a une structure modulaire, lui permettant éventuellement de s'adapter à toute extension et modification. Les connaissances spécifiques à l'application (courriels) sont générées automatiquement (niveau approfondi). En effet, le profil de l'utilisateur est calculé par analyse automatique du contenu qui permet de produire un ensemble de termes et de propriétés linguistiques le caractérisant.

Les résultats obtenus sur notre corpus semblent intéressants. Néanmoins, un ensemble d'extensions peuvent être envisagées :

- étendre l'étude sur d'autres types de courriels pour étendre et enrichir la liste des critères et tester l'adaptabilité de l'approche ;

- définir, pour la typologie du domaine d'application, une architecture générale et évolutive : l'architecture la plus adéquate est de structurer les différents types du domaine sous forme d'arborescence (hiérarchie) ouverte, appelée arbre de classification. Cette classification doit être la plus complète possible présentant les différents types, du type le plus général au plus spécifique. Cet arbre doit permettre l'héritage entre les types possédant des attributs communs et permettre d'ajouter (ou supprimer) des types jugés importants (ou non importants) par l'utilisateur ;
- intégrer, dans le processus de modélisation, les méthodes d'apprentissage ou de *classification non supervisée.* Ces méthodes sont dites méthodes de structuration. Elles sont caractérisées par la non disponibilité d'aucune autre information préalable que la description des exemples. Elles sont destinées à produire des groupements d'objets, selon un critère de similarité ou dissimilarité à partir d'une description sur ces objets (traits, caractéristiques, propriétés, etc.) ;
- traiter les anaphores pour diminuer les biais de calcul des cooccurrences ;
- etc.

A travers notre expérience, la conclusion que l'on peut évoquer est que l'apprentissage automatique est un passage obligé dans la conception et l'amélioration des performances d'un système de filtrage d'information dynamique et les méthodes linguistiques combinées aux méthodes statistiques semblent prometteuses pour avoir un filtrage efficace de l'information sur les réseaux de communication.

BIBLIOGRAPHIE

[1] AÏT-MOKHTAR (S.), CHANOD (J.P.), Xerox Incremental Parser (XIP). *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 72-79, 1997.

[2] AMINI (M.R.), Apprentissage automatique et recherche de l'information : application à l'extraction d'information de surface et au résumé de texte. *Thèse de doctorat*, Université de Paris 6, 2001.

[3] ANDROUTSOPOULOS (I.), KOUTSIAS (J.), CHANDRINOS (K.V.), PALIOURAS (G.), SPYROPOULOS (C.D.), An evaluation of naïve Bayesian anti-spam filtering. *Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000)*, Barcelona, Spain, pp. 9–17, 2000.

[4] APTÉ (C.), DAMERAU (F.), WEISS (S.), Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, **12**, n°3, pp. 233-251, 1994.

[5] BIBER (D.), Variation Across Speech and Writing. University Press, Cambridge, 1988.

[6] BRONCKART (J.P.), BAIN (D.), SCHNEUWLY (B.), DAVAUD (C.), PASQUIER (A.), Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse. *Lausanne: Delachaux & Niestlé,* 1985.

[7] BEN HAZEZ (S.), DESCLES (J.P.), MINEL (J.L.), Modèle d'exploration contextuelle pour l'analyse sémantique des textes. *TALN 2001,* Tours, pp. 73-82, 2001.

[8] BRILL (E.), A Simple Rule-based Part of Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, pp. 152-155, 1992.

[9] CAROPRESO (M.), MATWIN (S.), SEBASTIANI (F.), A learner-independent evaluation of the usefulness of statistical phrases for automatic text categorization, Hershey, US, pp. 78–102, 2001.

[10] CARRERAS (X.), MARQUEZ (L.), Boosting Trees for Anti-Spam Email Filtering. *Proceedings of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing,* 2001.

[11] CHANDRASEKAR (R.), SRINIVAS (B.), Using Syntactic Information in Document Filtering: A Comparative Study of Part-of-speech Tagging and super tagging. *Proceedings of the RIAO-97 Conference*, pp. 531-545, 1997.

[12] COLLINS (M. J.), A New Statistical Parser Based on Bigram Lexical Dependencies. *Proceedings of the 34th Annual Meeting of the ACL,* Santa Cruz, CA, 1996.

[13] COPECK (T.), BARKER (K.), DELISLE (S.), SZPAKOWICZ (S.), Automating the Measurement of Linguistic Features to Help Classify Texts as Technical. *TALN2000*, Lausanne, 2000.

[14] DAVALO (E.), NAIM (P.), Des Réseaux de Neurones. *Edition Eyrolles*, 1993.

[15] DESCLÉS (J.P.), CARTIER (E.), JACKIEWICZ (A.), MINEL (J.L.), Textual Processing and Contextual Exploration Method. *CONTEXT'97*, Rio de Janeiro, Brésil, pp. 189-197, 1997.

[16] DREYFUS (G.), MARTINEZ (J.M.), SAMUELIDES (M.), GORDON (M.B.), BADRAN (F.), THIRIA (S.), HERAULT (L.), Réseaux de neurones méthodologies et applications. *Eyrolles*, ISBN 2-212-11019-7, 2002.

[17] GARCIA (D.), Exploitation pour l'élaboration de requêtes de filtrage de texte, des connaissances causales détecté par COATIS. RIFRA'98 rencontre internationale sur l'extraction, le filtrage et le résumé automatique, pp. 44-54, 1998.

[18] HABERT (B.), ILLOUZ (G.), LAFON (P.), FLEURY (S.), FOLCH (H.), HEIDEN (S.), PREVOST (S.), Profilage de textes : cadre de travail et expérience. *JADT 2000: 5eme Journées Internationales d'Analyse Statistique des Données Textuelles,* 2000.

[19] JOACHIMS (T.), A probabilistic analysis of the Rocchio algorithm with tfidf for text categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pp. 143–151, 1997.

[20] JOACHIMS (T.), Text categorization with support vector machines: learning with many relevant features. *Proceeding of ECML-99, 16th European Conference on Machine Learning*, pp. 137–142, 1998.
[21] JUNKER (M.), ABECKER (A.), Exploiting Thesaurus Knowledge in Rule Induction for Text Classification. *Proceedings of the RANLP-97 Conference*, pp.202-207, 1997.
[22] LEWIS (D.D.), An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval,* Copenhagen, Denmark, pp. 35-50, 1992.
[23] LEWIS (D.D.), RINGUETTE (M.), Comparison of two learning algorithms for text categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval SDAIR'94,* 1994.
[24] MARCU (D.), From discourse structures to text summaries. *Workshop Intelligent Scalable Text Summarization,* Madrid, Spain, 1997.
[25] MINEL (J.L.), DESCLES (J.P.), CARTIER (E.), CRISPINO (G.), BEN HAZEZ (S.), JACKIEWICZ (A.), Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText. *Revue Technique et Science Informatique*, n° 3, 2001.
[26] MC CALLUM (A.), NIGAM (K.), A comparison of event models for naïve Bayes Text classification. L*earning for text categorization,* 1998.
[27] MILLER (G.), WordNet: An On-line Lexical Database. *International Journal of Lexicography,* 1990.
[28] Namer (F.), Flemm : Un analyseur Flexionnel du Français à base de règles. *Traitement automatique des langues pour la recherche d'information, revue T.A.L.*, **41**, n°2, (Jacquemin Ch. éd.), Paris, pp. 523-547*,* 2000.
[29] NOUALI (O.), Filtrage d'information textuelle sur les réseaux: une appeoche hybride, *thèse de doctorat*, université des sciences et technologie d'Alger, USTHB, 2004.
[30] ORASAN (C.), KRISHNAMURTHY (R.), A corpus-based investigation of junk emails. *Proceedings of LREC-2002,* Las Palmas, Spain, 2002.
[31] POIBEAU (T.), NAZARENKO (A.), L'extraction d'information, une nouvelle conception de la compréhension de texte? *TAL,* **40**, n°1-2, pp. 87-15, 1999.
[32] SAHAMI (M.), DUMAIS (S.), HECKERMAN (D.), HORVITZ (E.), A Bayesian approach to filtering junk e-mail. *In Learning for Text Categorization Papers from the AAAI Workshop,* Madison Wisconsin, pp. 55–62, 1998.
[33] SEBASTIANI (F.), A Tutorial on Automated Text Categorisation. *Proceedings of ASAI-99, 1st Argentinean Symposium on Artificial Intelligence,* 1999.
[34] YANG (Y.), PEDERSEN (J.O.), A comparative Study on Feature Selection in Text Categorization. *International Conference on Machine Learning ICML 1997,* Nashville, TN, USA, 1997.